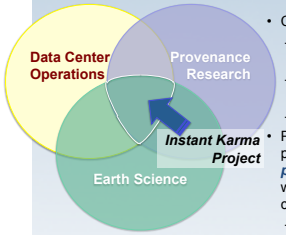
Provenance Collection and Display for the AMSR-E SIPS

Helen Conover, Bruce Beaumont, Ajinkya Kulkarni, Rahul Ramachandran, Kathryn Regner, Sara Graves, Dawn Conway
University of Alabama in Huntsville
http://provenance.itsc.uah.edu/

Approach



- Collaboration among
  - AMSR-E SIPS (MSFC Earth Science Office and UAHuntsville ITSC)
  - Provenance researchers at Indiana University's Data to Insight Center
  - AMSR-E Sea Ice science team (GSFC)
- Primary goal is to improve the collection, preservation, utility and dissemination of ***provenance and context information*** within the NASA Earth Science community
  - Using Karma provenance tool
  - AMSR-E standard product generation, with initial focus on sea ice

The Instant Karma project integrates Karma, a provenance collection and representation tool developed at Indiana University, into the AMSR-E Science Investigator-led Processing System (SIPS) production environment, managed jointly by NASA/MSFC and UAHuntsville. The AMSR-E SIPS generates Level 2 and Level 3 data products from AMSR-E observations. An initial focus on Sea Ice processing allows the project to engage the Sea Ice science team and user community in customizing provenance collection and display for NASA science data.

Provenance and Context Information

Data lineage (data inputs, software and hardware) plus additional ***contextual knowledge*** about science algorithms, instrument variations, etc.

Information already available, but scattered across multiple locations

- Processing system configuration
- Dataset and file level metadata
- Processing history information
- Quality assurance information
- Software documentation (e.g., algorithm theoretical basis documents, release notes)
- Data documentation (e.g., guide documents, README files)

Project collates and organizes information from multiple sources, available through the AMSR-E Provenance Browser

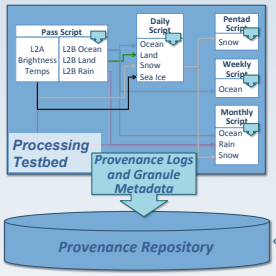
AMSR-E Science Use Cases

- ✓ Browse provenance graphs : convey rich information about final data granule details [Use case 1]
  - Spatial location, time of observation, algorithms employed, input data and ancillary files
  - Provenance bundle to include pointers to relevant documentation
- ✓ Answer "Something isn't right" question [Use case 1 variant]
  - E.g., did not receive data for several days so snow melt mask may be inaccurate.
- Compare two data granules [Use case 2]
  - Query system to get list of provenance differences (e.g., versions of software, number and versions of input files)
- ✓ General provenance graph for a given science process, e.g., Sea Ice processing [Use case 3]
  - Current algorithms and versions, nominal number and versions of input files, pointers to relevant documentation
- Embed provenance information as annotations in HDF files
  - Considering ISO "Lineage" model
  - Other NASA ES conventions?

Provenance Collection and Storage

AMSR-E SIPS processing workflows for all Level-2 and Level-3 products instrumented in the testbed environment.

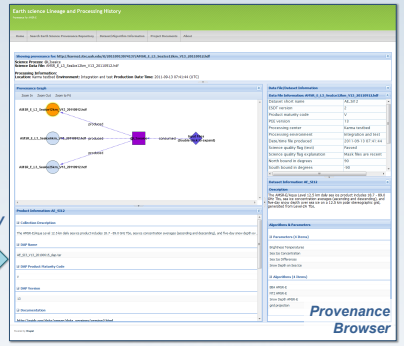
- Provenance information is captured in processing run log files
- Log files are parsed and imported into the provenance repository



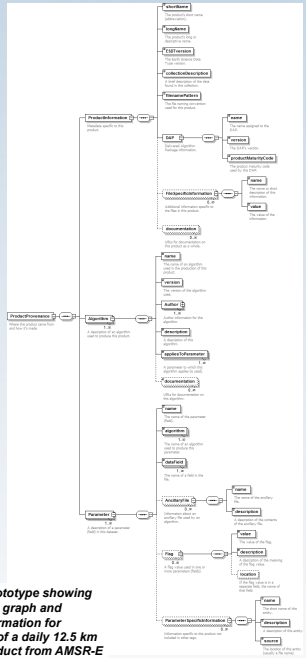
Query API

Defining and Collecting Science-Relevant Provenance and Context

- Harvesting granule information from ECS metadata
  - Also recording processing location associated with each data granule
- Working with AMSR-E Science Computing Facility to identify algorithm and data product information
  - Algorithm versions and descriptions
  - Parameters and data fields
  - Ancillary files
  - Flag values and explanations
  - Pointers to full documentation



Schema for high-level data product and algorithm context information



Browsing Provenance Information

- Interactive web application allows users to view the provenance graph for a specified data product
- Click on a node to display the full description of the product or process
- Trace full lineage of a data product by viewing the provenance information for each input file
- Access relevant information for the data product
  - Algorithm documentation and version information
  - README files
  - Product and inventory level metadata
- Uses query API to extract provenance graphs from the provenance repository.

Browser prototype showing provenance graph and related information for generation of a daily 12.5 km Sea Ice product from AMSR-E Brightness Temperatures.

Challenges

- Establishing communication and working relationships among diverse team
  - Different perspectives and vocabularies
  - Different software approaches
- Identifying "science-relevant" provenance among the many processing details that can be recorded
  - Reference implementation in Perl
- Engaging the science community
  - Need for compelling "science stories"

Status and Plans

- Currently able to collect provenance information for all AMSR-E standard products generated at SIPS-GHRC
  - Level-2B and Level-3 products
  - Implemented in provenance testbed
- Plan to implement provenance collection in AMSR-E SIPS before reprocessing to begin in February 2012
- Working with NSIDC DAAC on delivery of provenance information with data products
  - Content and format of information to provide
  - Include as annotation within the data granule, separate additional metadata file, or both

Potential for Reuse

- Software for provenance collection
  - Reference implementation in Perl
- Schema for high-level data product and algorithm context information
- Drupal profile for provenance storage and display, includes two new Drupal modules:
  - Provenance browser
  - Data product and algorithm context form

Acknowledgements: The Instant Karma project, funded under NASA ACCESS program, is a collaboration among NASA, UAHuntsville and Indiana University. The project team includes PI Michael Goodman (NASA/MSFC), Science Co-I Thorsten Markus (NASA/GSFC), Co-I's Helen Conover (UAHuntsville) and Beth Plale (IU) and their teams.

NASA
UAHuntsville THE UNIVERSITY OF ALABAMA IN HUNTSVILLE
DATA TO INSIGHT CENTER INDIANA UNIVERSITY PERVASIVE TECHNOLOGY INSTITUTE